DOI:
10.1109/TCOMM.2019.2938514

Document Version
Peer reviewed version

Calibrated Learning for Online Distributed Power Allocation in Small-Cell Networks

Xinruo Zhang, *Member, IEEE,* Mohammad Reza Nakhai, *Senior Member, IEEE,*
Gan Zheng, *Senior Member, IEEE,*
Sangarapillai Lambotharan, *Senior Member, IEEE,*
and Björn Ottersten, *Fellow, IEEE*

Abstract

This paper introduces a combined calibrated learning and bandit approach to online distributed power control in small cell networks operated under the same frequency bandwidth. Each small base station (SBS) is modelled as an intelligent agent who autonomously decides on its instantaneous transmit power level by predicting the transmitting policies of the other SBSs, namely the opponent SBSs, in the network, in real-time. The decision making process is based jointly on the past observations and the calibrated forecasts of the upcoming power allocation decisions of the opponent SBSs who inflict the dominant interferences on the agent. Furthermore, we integrate the proposed calibrated forecast process with a bandit policy to account for the wireless channel conditions unknown *a priori*, and develop an autonomous power allocation algorithm that is executable at individual SBSs to enhance the

Xinruo Zhang, Gan Zheng and Sangarapillai Lambotharan are with Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, LE11 3TU, U.K. (e-mail: {x.zhang, g.zheng, s.lambotharan}@lboro.ac.uk).

Mohammad Reza Nakhai is with Centre for Telecommunications Research, King's College London, WC2B 4BG, U.K. (e-mail: reza.nakhai@kcl.ac.uk).

Björn Ottersten is with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City L-1855, Luxembourg. (e-mail: bjorn.ottersten@uni.lu).

accuracy of the autonomous decision making. We evaluate the performance of the proposed algorithm in cases of maximizing the long-term sum-rate, the overall energy efficiency and the average minimum achievable data rate. Numerical simulation results demonstrate that the proposed design outperforms the benchmark scheme with limited amount of information exchange and rapidly approaches towards the optimal centralized solution for all case studies.

Index Terms

small cell; distributed power control; online learning; calibration

I. INTRODUCTION

Cellular networks have experienced an explosive increase in the number of mobile subscribers and wireless devices in the last decade [1]. To this end, network densification has become a general trend for future wireless networks. Hyper-dense deployment of small base stations (SBSs) that provide short-range, low-power and low-cost transmission in addition to the existing macrocell cellular networks, is considered a promising technique to meet the requirements of the mounting growth of mobile data traffic and massive connectivity [2]. To fully exploit the potentials of small cells, full frequency reuse among small cells is necessary [3]. However, such densification in cellular networks with limited licensed spectrum will result in increasing intercell interference (ICI), as increasing transmit power of a transmitter for the improvement of its own capacity may degrade other links it interferes with. Transmit power control and interference management mechanisms have been investigated for long to ensure the system-level performance. Cloud radio access networks that allow centralized radio resource coordination across multiple cells, have been widely examined in the literature for ICI management and resource allocation [4]. Although having a central node for signalling coordination is one option for power control and interference management, it is not always possible or preferable in many scenarios to have a central coordination node available who can collect the necessary information, e.g., channel state information (CSI) and user requirements of the entire network, and then perform radio resource

allocation. In practice, multiple small cell network operators may access shared spectrum [5], [6]. For instance, Citizen Broadband Radio Service [7] allows shared spectrum use for 3.5 GHz band via a three-tier access model [5], where the users in the lowest tier can access spectrum unused by the higher tier users. From the network operators' perspectives, it may not be practical and preferable to have a central coordination node for the purpose of inter-operator power control and interference management, as the operators are competitors to each other and currently no generally acknowledged mechanism is available for the operator-level coordination. Furthermore, considering the density and randomness of SBS deployment, the backhaul link between the central node and the SBSs tends to be wireless, which, has limited capacity and can be vulnerable to dynamic changes in the environment [3]. Besides, relying on the central control node in hyper-dense small cell networks to collect CSI of the entire network and perform system-level power control can be computationally complex and may fail to provide quick response to the channel and traffic variations [8]. Consequently, the SBSs are proposed to manage radio resource distributively with limited and time-insensitive information exchange over the X2 interface with each other [8], which necessitates the distributed design of SBSs that can be self-organized and coordinated autonomously [9].

A. Related Works

The distributed resource allocation problem for collaborative base stations (BSs) has been widely studied in the literature [10]–[24]. The distributed power control for max-min signal-to-interference-plus-noise ratio (SINR) problem is studied in [10] in a simple scenario with only 2 SBSs and 2 users. The authors in [11] assume that the ICI is an ergodic Gaussian process and the induced statistical average ICI power can be estimated by the users and reported to the BSs. Other conventional literature either models the distributed resource allocation problem as a noncooperative game and solves it via iterative processes [12]–[14], or uses iterative inter-BS fronthaul information exchange such as the subgradient method in [15]–[19] and the alternating

direction method of multipliers (ADMM) technique in [19], [20] to schedule transmit power and manage ICI among BSs. Focusing on uplink heterogeneous networks (HetNets), the authors in [13] adopt the Debreu-type noncooperative game and introduce a distributed power allocation algorithm based on an iterative water-filling best response process. The authors in [14] model the interference problem between a macrocell BS and multiple SBSs as a Stackelberg game, and claim that the macrocell BS can make optimal power allocation decisions based on its prediction of the reactions of SBSs. However, the leader-follower game such as the Stackelberg game that requests sequential moves may not be suitable for the practical scenario where BSs perform simultaneous transmissions. The authors in [15]–[20], on the other hand, address the power allocation problem via inter-BS communications and alternate between independently solving a convex power optimization problem at individual BSs and circulating key intercell coupling parameters among BSs. The aforementioned designs, nevertheless, highly rely on the tractable mathematical models, which may not be available in practical scenarios. Furthermore, they assume that the channel remains invariant until the iterations are completed or the convergence is achieved. Such assumption is not very practical as the resulting signalling overhead depending upon the number of iterations may exceed the short channel coherence time.

Recently, various machine learning based approaches for resource allocation have been developed in wireless communications [21]–[26]. The authors in [21] adopt calibration based no-regret bandit learning approach to maximize the transmission rate in D2D network with fixed transmit power. In [25], the authors propose a centralized supervised learning approach to train a deep neural network to approximate the resource allocation algorithms. The supervised learning, however, depends highly on the accuracy of the system model and may require a new set of training data when key parameters change. The authors in [26] propose a learning to optimize framework to accelerate the branch-and-bound algorithm for centralized interference channel binary power control. Considering the quality of service and user fairness, the authors in [22] propose a cooperative Q-learning based power allocation mechanism in HetNets. However, all of

the channel information is assumed to be known at BSs and sharing Q-values to reduce search time of the agents may raise additional signalling overhead issues. The authors in [23], [24] introduce deep Q-learning approaches to power allocation problem in single user and multi-user wireless networks, respectively. However, prior to distributed execution at individual BSs, their proposed approaches require the deep Q network to be trained off-line by a centralized network trainer via experience replay from the data set pool.

B. Contribution

This paper focuses on the design of an online distributed[1] mechanism for intelligent power allocation in a small-cell network operating in a shared frequency bandwidth over a long time horizon. An ϵ-calibration based bandit approach is developed for each individual SBS to asymptotically and distributively calibrate its forecast on the possible transmit power levels of the other SBSs, and react with the best power allocation decision based on the predicted results as well as the past observations. The contributions of this paper are summarized as follows:

- In contrast to the existing distributed power allocation designs that iteratively exchange information among SBSs within a channel coherence time [15]–[20], our work considers a more realistic scenario, where SBSs are autonomous decision makers and they exchange only the past power information at the end of each time instance.
- The proposed algorithm requires neither prior knowledge of the systems nor the statistical distribution of the wireless channel conditions. Unlike deep neural network based power allocation designs [23]–[25] that need to be off-line trained by a centralized network trainer based on the collected experiences, the proposed distributed design requires no centralized coordination or training, and performs online calibrated learning during the operation.

[1]The distributed framework in the strict sense relies only on local statistics available at each end. Throughout this paper, it is considered in the broad sense, where there is no central coordination unit, and the computational tasks as well as the decision making processes have been shifted to the individual SBSs at the cost of limited information exchange among SBSs.

C. Organization and Notations

The rest of this paper is organized as follows. Section II introduces the system model and formulates a long-term power control problem. In section III, the ϵ-calibration will be first introduced, followed by the proposed calibration based online learning algorithm. Numerical simulation results are analyzed in section IV. Finally, section V concludes the paper.

Notations: w and $\mathbf{w}$, respectively, indicate a scalar w and a vector $\mathbf{w}$. The notations $\mathbb{B}^n$, $\mathbb{R}^{nm}$ and $\mathbf{0}$ denote, respectively, the binary space of n-dimension vectors, the sets of m dimensional real vectors with components in $\mathbb{R}^n$, and an all-zero matrix with appropriate dimensionality. $[\mathbf{w}_1|\mathbf{w}_2]$ indicates the concatenation of vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. $\mathbf{1}_{\{.\}}$ is an indicator function that returns one if $\{.\}$ holds true and zero otherwise.

II. SYSTEM MODEL AND PROBLEM FORMULATION

This paper considers a distributed downlink small cell network consisting of N_b SBSs, indexed by $\mathcal{L}_b = \{1, \cdots, N_b\}$, that serve their own N_u users over a shared spectrum. Let the time horizon T be divided into discrete time slots with slot duration of τ, and indexed as $\mathcal{T} = \{1, \cdots, T\}$. Each time slot corresponds to a channel coherence time and the channel is assumed to vary across time slots but remains invariant within each time slot. The time-division multiple access technology is employed to ensure user fairness via serving the users within the same small cell one after the other in their respective time slots. For the sake of notational simplicity, let us assume that the individual SBSs have A discrete power levels and denote by $\mathcal{A} = \{1, \cdots, A\}$ and $\mathcal{E} = \{E_1, \cdots, E_A\}$, respectively, the indexes and the set of discrete transmit power of the SBSs. Note that although we consider the same $\mathcal{E}$ for each SBS, the proposed approach can be easily adapted to the scenarios where the power region and/or the number of divisions of transmit power of each SBS are distinct. Each individual SBS is fully synchronized and has to make its own decision on selecting a transmit power from the set $\mathcal{E}$, independently and simultaneously, to serve its scheduled user during time slot $t, t \in \mathcal{T}$. Furthermore, the SBSs can communicate their

TABLE I

NOTATION

Symbol	Definition
$\mathcal{L}_b = \{1, \cdots, N_b\}$	Index set of N_b SBSs
$b' \in \mathcal{L}_b \backslash b$	SBS b', $b' \in \mathcal{L}_b, b' \neq b$
$\mathcal{T} = \{1, \cdots, T\}$	Index set of discrete time slots with slot duration τ
$\mathcal{A} = \{1, \cdots, A\}$	Index set of A discrete transmit power levels (actions) of each SBS
$\mathcal{E} = \{E_1, \cdots, E_A\}$	Set of A discrete transmit power of eahc SBS
$P_b^t \in \mathcal{E}$	Transmit power from SBS b to its scheduled user at time t
Ψ_{bb}^t	Channel gain between SBS b and its scheduled user at time t
R_b^t	Downlink data rate for scheduled user of SBS b at time t
$a_b^t \in \mathcal{A}$	Actual transmit power level (action) of SBS b at time t
$\boldsymbol{\delta}_{a_{b'}^t} \in \mathbb{B}^A$	Dirac probability distribution on SBS b''s actual action $a_{b'}^t$ at time t
$\mathcal{P} = \triangle(\mathcal{A}) \subset \mathbb{R}^A$	Set of candidate probability distributions over A possible outcomes, i.e., A possible transmit power levels, for calibration
ϵ^t	ϵ-calibration at time t for Algorithm 1
N_ϵ	Number of candidate probability values for ϵ-calibration
$\hat{\mathcal{P}} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\} \subset \mathbb{R}^A$	Set of N_ϵ candidate probability values for ϵ-calibration. Each element $\mathbf{p}_k = [p_{k,1}, \ldots, p_{k,A}] \in \mathbb{R}^A$, $k \in \{1, \ldots, N_\epsilon\}$ is the probabilities over all of the A possible transmit power levels (outcomes) and $\sum_{a=1}^{A} p_{k,a} = 1$.
$\mathbf{p}_{bb'}^t = [p_{bb',1}^t, \ldots, p_{bb',A}^t] \in \hat{\mathcal{P}}$	SBS b's forecasted probabilities over all possible outcomes of SBS b' at time t
$\mathcal{C} \subset \mathbb{R}^{AN_\epsilon}$	Closed convex target set for approachability theorem
$\mathbf{m}(k, a) \in \mathbb{R}^{AN_\epsilon}$	Vector-valued regret between prediction k of the forecaster and action a of the opponent at a certain time slot, $\forall k \in \{1, \ldots, N_\epsilon\}, a \in \mathcal{A}$
$\overline{\mathbf{m}}_{bb'}^T \in \mathbb{R}^{AN_\epsilon}$	Average vector-valued regrets between SBS b's predictions on SBS b' and SBS b''s true actions up to time T
$\mathbf{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^T) \in \mathbb{R}^{AN_\epsilon}$	Projection of $\overline{\mathbf{m}}_{bb'}^T$ in l_2-norm onto $\mathcal{C}$
$\boldsymbol{\psi}_{bb'}^t = [\psi_{bb',1}^t, \cdots, \psi_{bb',N_\epsilon}^t] \in \mathbb{R}^{N_\epsilon}$	Prediction distribution over $\hat{\mathcal{P}} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ at SBS b for SBS b' at time t
$\hat{\boldsymbol{\mu}}_b^t = [\hat{\mu}_{b,1}, \cdots, \hat{\mu}_{b,A}] \in \mathbb{R}^A$	Estimated mean reward vector of SBS b at time t
$\bar{\boldsymbol{\mu}}_{bb'}^t = [\bar{\mu}_{bb',1}, \cdots, \bar{\mu}_{bb',A}] \in \mathbb{R}^A$	Discounted mean reward vector at SBS b for SBS b' at time t
β^t	Discount factor at time t
γ_t	Adaptive exploration-exploitation trade-off at time t

past information such as the actual past transmitted power levels (actions) with each other via capacity-constrained inter-SBS communication links at the end of each time slot. The notations in this paper are listed in Table I.

A. Downlink Transmission Model

Let the channel gain between SBS b, $b \in \mathcal{L}_b$ and its scheduled user at the t-th time slot, $t \in \mathcal{T}$, be denoted by Ψ_{bb}^t. We adopt a block fading channel model [27] for the downlink channel gain, as

$$\Psi_{bb}^t = |h_{bb}^t|^2 \alpha_{bb}^t, \tag{1}$$

where $\alpha_{bb}^t \geq 0$ captures all large-scale fading effects including path loss and log-normal shadowing between SBS b and its scheduled user at time t, and h_{bb}^t is the small-scale fast fading component. We adopt Jakes' model [27] to model the channel variation over the period τ and denote the Rayleigh distributed small-scale fading component as a first-order Gauss-Markov process, as

$$h_{bb}^t = \rho h_{bb}^{t-1} + \sqrt{1-\rho^2} d_{bb}^t, \tag{2}$$

where $h_{bb}^0 \in \mathbb{CN}(0,1)$ is the circularly symmetric complex Gaussian random variable with unit variance, and $d_{bb}^t \in \mathbb{CN}(0,1)$ is the independent and identically distributed (i.i.d) channel discrepancy term. The coefficient $0 \leq \rho < 1$ quantifies the channel correlation between two consecutive time slots h_{bb}^t and h_{bb}^{t-1}, and is given by $\rho = J_0(2\pi f_d \tau)$, where $J_0(\centerdot)$ is the zero-order Bessel function of the first kind, f_d denotes the maximum Doppler frequency and τ is the duration of each time slot. Let us denote by $P_b^t \in \mathcal{E}$ the transmit power from SBS b to its scheduled user at time t. Then, the SINR for the scheduled user served by SBS $b,\ b \in \mathcal{L}_b$ at time slot $t,\ t \in \mathcal{T}$, is given by

$$\text{SINR}_b^t = \frac{P_b^t \Psi_{bb}^t}{\sum\limits_{b' \in \mathcal{L}_b \setminus b} P_{b'}^t \Psi_{b'b}^t + \sigma_b^2}, \tag{3}$$

where the numerator of (3) denotes the desired signal power for the served user, the terms in the denominator denote, respectively, the ICI caused by all other SBSs and the additive white Gaussian noise (AWGN) with variance of σ_b^2 at the user. The cross-link channel gain $\Psi_{b'b}^t$ can be fed back by the user to its associated SBS. Then, with the normalized bandwidth, the instantaneous downlink data rate for the scheduled user served by the b-th SBS, $b \in \mathcal{L}_b$, at time slot t, $t \in \mathcal{T}$, can be expressed as

$$R_b^t = \log_2(1 + \text{SINR}_b^t). \tag{4}$$

B. Problem Formulation

In the considered distributed small cell network, it is assumed that the SBSs are autonomous decision makers, i.e., they have to make their decisions independently and distributively without knowing the resulting impacts, such as interference, on one another. It is evident that without a proper power control mechanism, the presence of interference, coupling with the decisions of individual SBSs, leads to a conflict of interest: each individual SBS attempts to transmit at a power level that maximizes its own utility, while the interference incurred by doing so may degrade the performance of other SBSs. Thus, we formulate the transmit power control problem for distributed small cell networks as a long-term system-level reward maximization problem to optimize the power allocation decisions of all SBSs, as

$$\begin{aligned} &\max_{\{P_b^t\}} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{U}(\{R_b^t\}) \right\} \\ &\text{s.t.} \quad P_b^t \le P_b^{\max}, \ \forall b \in \mathcal{L}_b, \end{aligned} \tag{5}$$

where $P_b^{\max}$ denotes the maximum transmit power of SBS b. $\mathbf{U}(\{R_b^t\})$ is a utility function of $\{R_b^t\}_{b\in\mathcal{L}_b}$ that indicates the system-level performance over all SBSs, and can be adapted to various system objectives. Note that in the considered distributed scenario, the SBSs have to make decisions simultaneously and independently, and only the outdated power information of the other SBSs is acquirable at each SBS. Based upon this information, the individual SBSs

attempt to estimate the possible transmit power of the others at current time t, and react with the best response based on their estimations. Hence, the key issue to be addressed for each individual SBS is the reliable prediction of the transmit power decisions of the other SBSs. In the next section, a calibrated learning based online distributed power allocation algorithm will be proposed for each SBS $b, b \in \mathcal{L}_b$, to gradually improve its accuracy in predicting the transmit power of the other SBSs, i.e., $\{P_{b'}^t\}_{b' \in \mathcal{L}_b \backslash b}$, and select its most appropriate transmit power, i.e., P_b^t, such that the system-level performance can be optimized in the long run.

III. CALIBRATION BASED ONLINE LEARNING ALGORITHM

In the sequel, the ϵ-calibrated forecaster [28] will be firstly introduced for each individual SBS to gradually improve its accuracy in predicting the transmit power decisions of the other SBSs, namely the opponent SBSs. It will be followed by the proposed calibrated learning based online distributed power allocation algorithm that allows each individual SBS to select the most appropriate transmit power level on the basis of the possible choices as well as the corresponding impacts of the other SBSs, so as to asymptotically maximize the average reward in the long run.

A. ϵ-Calibrated Forecaster for the Opponents' Joint Action

Calibration is a central tool in game theory and in online learning for prediction and forecasting [28]. The definition of calibration can be presented as follows. Consider a finite set of possible outcomes. Observing the environment, a forecaster assigns subjective probabilities (forecasts) over the possible outcomes to future events. The sequence of forecasts is called calibrated if, with time, the forecasted probabilities converge to the observed long-term empirical (true) probabilities of the outcomes.

In our considered scenario, the finite set of possible outcomes is a finite set $\mathcal{A}$ of transmit power levels with cardinality of A. Each individual SBS $b, b \in \mathcal{L}_b$, can be regarded as a forecaster that attempts to predict the probabilities over the possible transmit power levels of its opponents,

i.e., SBS $b',\ b' \in \mathcal{L}_b \backslash b$. Let us denote by $\mathcal{P} = \triangle(\mathcal{A}) \subset \mathbb{R}^A$ the set of pre-defined candidate probability distributions over A outcomes and denote by $\mathbf{p}_{bb'}^t = [p_{bb',1}^t, \ldots, p_{bb',A}^t] \in \mathcal{P}$ the SBS b's forecasted probabilities over all of the A possible outcomes of SBS b' at time t, where $\sum_{a=1}^{A} p_{bb',a}^t = 1$. At each time slot $t,\ t \in \mathcal{T}$, SBS b outputs its forecasted probabilities $\mathbf{p}_{bb'}^t \in \mathbb{R}^A$ over all possible transmit power levels of SBS b', whilst SBS b' selects an outcome, i.e., a transmit power level $a_{b'}^t \in \mathcal{A}$. Then, the calibrated forecaster [28] can be defined as follows: for $\forall \epsilon > 0,\ \forall \hat{\mathbf{p}} \in \mathcal{P}$, almost surely,

$$\lim_{T\to\infty} \| \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\{\|\mathbf{p}_{bb'}^t - \hat{\mathbf{p}}\| \leq \epsilon\}} (\mathbf{p}_{bb'}^t - \boldsymbol{\delta}_{a_{b'}^t}) \| = 0, \tag{6}$$

where $\boldsymbol{\delta}_{a_{b'}^t} \in \mathbb{B}^A$ is the dirac probability distribution on SBS b''s outcome $a_{b'}^t$ at time t, and $\mathbf{1}_{\{\|\mathbf{p}_{bb'}^t - \hat{\mathbf{p}}\| \leq \epsilon\}}$ is an indicator function that returns one if $\|\mathbf{p}_{bb'}^t - \hat{\mathbf{p}}\| \leq \epsilon$ holds true and zero otherwise.

ϵ-calibration, on the other hand, can be regarded as a relaxed version of calibration in (6) by reducing $\mathcal{P}$ to a finite set of N_ϵ pre-defined candidate probability values, i.e., $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$, for a given ϵ. Given ϵ, the forecaster b considers some finite covering of $\mathcal{P}$ by N_ϵ balls of radius ϵ, where the centers of the balls in the covering is denoted as $\hat{\mathcal{P}} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ [29], and SBS b only outputs its forecasted probabilities of opponent b' from a finite set $\hat{\mathcal{P}}$ of candidate probability values, i.e., $\mathbf{p}_{bb'}^t \in \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$. Then, the ϵ-calibrated forecaster [29] can be defined as:

$$\limsup_{T\to\infty} \sum_{k=1}^{N_\epsilon} \| \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\{K^t = k\}} (\hat{\mathbf{p}}_k - \boldsymbol{\delta}_{a_{b'}^t}) \| \leq \epsilon, \tag{7}$$

where K^t denotes the index in $\{1, \ldots, N_\epsilon\}$ such that $\mathbf{p}_{bb'}^t = \hat{\mathbf{p}}_{K^t}$. As demonstrated in [29], calibration is a consequence of approachability, i.e., the existence of ϵ-calibrated forecaster b in (7) against its opponent b' is equivalent to the approachability of some closed convex target set $\mathcal{C}$. Following a similar procedure as in [29], for a given ϵ, let us define the target set $\mathcal{C}$ as follows

$$\mathcal{C} = \{\underline{\mathbf{x}} \in \mathbb{R}^{AN_\epsilon} : \sum_{k=1}^{N_\epsilon} \|\mathbf{x}_k\| \leq \epsilon\}, \tag{8}$$

where $\underline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_\epsilon})$ and $\mathbf{x}_k \in \mathbb{R}^A, \forall k \in \{1, \ldots, N_\epsilon\}$. Let us define the vector-valued regret between the forecaster's predicted probabilities $\hat{\mathbf{p}}_k$ and the dirac probability distribution on opponent's action a at a certain time slot as $\forall k \in \{1, \ldots, N_\epsilon\}, a \in \mathcal{A}$,

$$\mathbf{m}(k, a) = (\mathbf{0}, \ldots, \mathbf{0}, \hat{\mathbf{p}}_k - \boldsymbol{\delta}_a, \mathbf{0}, \ldots, \mathbf{0}), \tag{9}$$

where $\mathbf{m}(k, a) \in \mathbb{R}^{AN_\epsilon}$ contains one non-zero vector element of $\mathbb{R}^A$ located at the k-th position and $N_\epsilon - 1$ zero vector elements elsewhere. According to Blackwell's approachability theorem [30], $\mathcal{C}$ is approachable if and only if $\forall a \in \mathcal{A},\ \exists k \in \{1, \ldots, N_\epsilon\}$, $\mathbf{m}(k, a) \in \mathcal{C}$. Then, the ϵ-calibration in (7) can be expressed in the following way: there exists a strategy for forecaster b that ensures that for all strategies (actions) of its opponent b', the average of the vector-valued regrets, i.e.,

$$\overline{\mathbf{m}}_{bb'}^T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{m}(K^t, a_{b'}^t) = \frac{1}{T}(\sum_{t=1}^{T} \mathbf{1}_{\{K^t=1\}}(\hat{\mathbf{p}}_1 - \boldsymbol{\delta}_{a_{b'}^t}), \ldots, \sum_{t=1}^{T} \mathbf{1}_{\{K^t=N_\epsilon\}}(\hat{\mathbf{p}}_{N_\epsilon} - \boldsymbol{\delta}_{a_{b'}^t})) \tag{10}$$

converges to the set $\mathcal{C}$ almost surely [29]. It has been proved in [29] that there exists an ϵ-calibrated forecaster b, who selects at each time t an aforementioned strategy $\mathbf{p}_{bb'}^t \in \{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ according to the optimal prediction distribution $\boldsymbol{\psi}_{bb'}^{t*} = [\psi_{bb',1}^{t*}, \cdots, \psi_{bb',N_\epsilon}^{t*}]$ over $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$, such that for $\forall a \in \mathcal{A}$ of its opponent b',

$$(\overline{\mathbf{m}}_{bb'}^{t-1} - \boldsymbol{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^{t-1})) \cdot (\mathbf{m}(\boldsymbol{\psi}_{bb'}^{t*}, a) - \boldsymbol{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^{t-1})) \leq 0, \tag{11}$$

where $\cdot$ denotes the inner product, $\boldsymbol{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^{t-1})$ represents the projection of $\overline{\mathbf{m}}_{bb'}^{t-1}$ in l_2-norm onto $\mathcal{C}$, and $\mathbf{m}(\boldsymbol{\psi}_{bb'}^{t*}, a) = \sum_{k=1}^{N_\epsilon} \psi_{bb',k}^t \mathbf{m}(k, a)$ is the weighted sum of vector-valued regret of forecaster b at time t. In other words, ϵ-calibration ensures that with time, the average vector-valued regrets of forecaster b lies in the convex set $\mathcal{C}$ almost surely. As such $\boldsymbol{\psi}_{bb'}^{t*}$ indeed exists [29], it suffices

to solve the following problem for finding the optimal prediction distribution $\boldsymbol{\psi}_{bb'}^{t*}$, as

$$\begin{aligned}
&\underset{\boldsymbol{\psi}_{bb'}^{t}}{\operatorname{argmin}} \max_{a\in\mathcal{A}} (\overline{\mathbf{m}}_{bb'}^{t-1} - \boldsymbol{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^{t-1})) \cdot \mathbf{m}(\boldsymbol{\psi}_{bb'}^{t}, a) = \\
&\underset{\boldsymbol{\psi}_{bb'}^{t}}{\operatorname{argmin}} \max_{a\in\mathcal{A}} \sum_{k=1}^{N_\epsilon} \psi_{bb',k}^{t} (\overline{\mathbf{m}}_{bb'}^{t-1} - \boldsymbol{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^{t-1})) \cdot \mathbf{m}(k, a) \\
&\text{subject to } \sum_{k=1}^{N_\epsilon} \psi_{bb',k}^{t} = 1,
\end{aligned} \tag{12}$$

which can either be solved exactly via linear programming, or be approximated via the multiplicative weights algorithm [31]. Recall that each SBS b acts as a forecaster, the other SBSs $b',\ b' \in \mathcal{L}_b \backslash b$, correspond to the opponents, and the actions of the opponents can be regarded as the other SBSs transmitting to their respective scheduled users at certain discrete power levels. The procedure of the ϵ-calibrated forecaster at each SBS b to predict the possible actions of the other SBSs are summarized in Algorithm 1.

B. Calibrated Learning Algorithm for Power Allocation

Since the wireless channel conditions vary across time slots and are unknown in advance, in order to maximize the long-term system-level performance over all SBSs in such a distributed scenario, the impacts of these uncertain factors on the objective of the problem in (5) need to be learned over time. Here we propose a distributed bandit approach to take these impacts into account and to enhance the the autonomous decision making process.

Let us consider a bandit problem that models a system of A actions whose expected rewards are i.i.d over time with unknown means. The objective is to maximize the accumulated reward over time via exploring the environment to find profitable actions, while exploiting current knowledge to make the empirically best decisions among a set of actions [32]. The online distributed power control problem investigated in this paper can be regarded as a bandit problem, where each individual SBS acts as an agent and a forecaster, A transmit power levels correspond to A actions, and the instantaneous reward of individual transmit power levels chosen by SBS $b, b \in \mathcal{L}_b$, for

Algorithm 1 *ϵ-calibrated forecaster at SBS b to predict its opponents' actions*

1: **Input**: current time-slot t, ϵ^t, discount factor β^t, actual transmit power level chosen by the opponent SBSs at previous time slot $\{a_{b'}^{t-1}\}$.

2: **If** $t = 1$

Set $\boldsymbol{\psi}_{bb'}^{t*} = [\frac{1}{N_\epsilon}, \cdots, \frac{1}{N_\epsilon}],\ \forall b' \in \mathcal{L}_b \backslash b$.

3: **Else**

4: Update the discounted average vector-valued regrets up to time $t-1$ as per (10) for each opponent b', as $\overline{\mathbf{m}}_{bb'}^{t-1} = \frac{1}{t-1} \sum_{t'=1}^{t-1} \mathbf{m}(K^{t'}, a_{b'}^{t'}) \prod_{i=t'}^{t-1} \beta^i,\ \forall b' \in \mathcal{L}_b \backslash b$.

5: Calculate the projection $\mathbf{\Gamma}_{\mathcal{C}}(\overline{\mathbf{m}}_{bb'}^{t-1}),\ \forall b' \in \mathcal{L}_b \backslash b$, by obtaining the optimal solution $\underline{\mathbf{x}}^* \in \mathbb{R}^{AN_\epsilon}$ of the following problem

$$\min_{\underline{\mathbf{x}}} ||\underline{\mathbf{x}} - \overline{\mathbf{m}}_{bb'}^{t-1}||_2^2$$
$$\text{s.t. } x_k \geq 0,\ \ 1 \leq k \leq AN_\epsilon,\ \ \textstyle\sum_{k=1}^{AN_\epsilon} x_k \leq \epsilon^t.$$

6: Optimize the prediction distributions $\{\boldsymbol{\psi}_{bb'}^t\}$ by solving problem (12), such that for $\forall a \in \mathcal{A}$, (11) is satisfied [29].

7: **End If**

8: Select the forecast probabilities $\mathbf{p}_{bb'}^t$ according to the optimal prediction distribution $\boldsymbol{\psi}_{bb'}^{t*}$ over $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$, $\forall b' \in \mathcal{L}_b \backslash b$.

9: **Output**: $\{\mathbf{p}_{bb'}^t\}$

serving its scheduled user at time t, can be defined as the normalized data rate at the t-th time slot, i.e., R_b^t.

The proposed distributed calibrated learning based power allocation algorithm is governed by a trade-off between exploring different actions that might yield a better accumulated reward in the presence of wireless channel random dynamism, and exploiting the best action so far that maximizes the accumulated reward at the individual SBSs. The details of the proposed

Algorithm 2 *The Main Distributed Calibration based Learning Algorithm at SBS b*

Initialize: time slot count $t = 1$; total number of time slots T, estimated mean reward vector $\hat{\boldsymbol{\mu}}_b^t = [\hat{\mu}_{b,1}, \cdots, \hat{\mu}_{b,A}] = \mathbf{0}$, discounted mean reward vector $\bar{\boldsymbol{\mu}}_{bb'}^t = [\bar{\mu}_{bb',1}, \cdots, \bar{\mu}_{bb',A}] = \mathbf{0}$, ϵ^t, discount factor β^t, adaptive exploration-exploitation trade-off γ_t.

while $t \neq T$ **do**

with the probability of γ_t: **Exploration Stage**

-Transmit to its scheduled user at a random power level.

with the probability of $1 - \gamma_t$: **Exploitation Stage**

-Receive the forecasts of probabilities over possible actions of SBS b', $b' \in \mathcal{L}_b \backslash b$, for current time slot t, i.e., $\{\mathbf{p}_{bb'}^t = \{p_{bb',a'}^t\}_{a' \in \mathcal{A}}\}$, from Algorithm 1.

-Calculate the estimated mean reward as $\hat{\mu}_{b,a}^t = \sum_{\mathbf{s}_b \in \mathcal{S}_b} p_{b,\mathbf{s}_b}^t \hat{R}_b^t(a, \mathbf{s}_b), \forall a \in \mathcal{A}$, where $\mathcal{S}_b = \{\{a_{b'}\}_{b' \in \mathcal{L}_b \backslash b} \, | a_{b'} \in \mathcal{A}\}$ contains all of the $|\mathcal{S}_b| = A^{N_b - 1}$ action combinations of the opponent SBSs and $p_{b,\mathbf{s}_b}^t = \prod_{b' \in \mathcal{L}_b \backslash b} p_{bb',\mathbf{s}_b(b')}^t$ is the joint forecast probability. $\hat{R}_b^t(a, \mathbf{s}_b)$ is the estimated data rate when SBS b selects action a and the other SBSs pick a joint action $\mathbf{s}_b \in \mathcal{S}_b$, and can be computed by substituting $P_b^t = E_a$ and $\{P_{b'}^t\} = E_{\mathbf{s}_b}$ into (4).

-Transmit at the power level associated with the highest overall reward, as $a_b^t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \mathbf{U}(\hat{\boldsymbol{\mu}}_b^t, \{\bar{\boldsymbol{\mu}}_{b'b}^{t-1}\})$, where $\mathbf{U}(\cdot)$ is a function of its estimated mean reward $\hat{\boldsymbol{\mu}}_b^t$ and the discounted mean rewards $\{\bar{\boldsymbol{\mu}}_{b'b}^{t-1}\}$, which corresponds to the system objective.

Observe the true instantaneous reward R_b^t associated to the selected transmit power level.

Calculate the discounted mean reward $\bar{\boldsymbol{\mu}}_{bb'}^t$, as

$$\bar{\mu}_{bb',a}^t = \frac{\sum_{t'=1}^{t-1} \mathbf{1}_{a_{b'}^{t'}=a} R_b^{t'} \prod_{i=t'}^{t-1} \beta^i}{\sum_{t'=1}^{t-1} \mathbf{1}_{\{a_{b'}^{t'}=a\}} \prod_{i=t'}^{t-1} \beta^i}, \ \forall a \in \mathcal{A}, b' \in \mathcal{L}_b \backslash b.$$

Share its actual power decision, i.e., a_b^t, and the discounted mean reward information $\bar{\boldsymbol{\mu}}_{bb'}^t$ with SBS b', $\forall b' \in \mathcal{L}_b \backslash b$.

increment $t = t + 1$.

end while

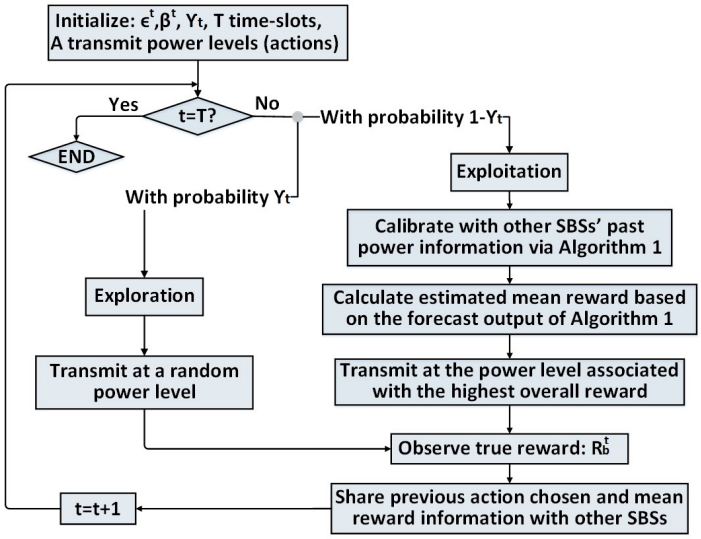


Fig. 1. Flowchart of the proposed Algorithm 2 at the individual SBSs.

algorithm to be executed at the individual SBSs are described in Algorithm 2 and Fig. 1, where the individual time slots can either be allocated for exploration with the probability of γ_t, or for exploitation with the probability of $1-\gamma_t$, as explained below:

- Exploration: With the probability of γ_t, a perturbation procedure is applied to explore the environment, e.g., the time-varying wireless channel conditions and the possible choice of the other SBSs, by transmitting at a random power level. Then, each SBS b will receive the feedback of quality of service, i.e., the associated instantaneous reward R_b^t, from its scheduled user for future action selection.
- Exploitation: With the probability of $1-\gamma_t$, each SBS b will transmit at the power level that yields the highest overall reward so far, based on the predicted results, i.e., $\mathbf{p}_{bb'}^t = \{p_{bb',a'}^t\}_{a'\in\mathcal{A}}, \forall b' \in \mathcal{L}_b \backslash b$, from Algorithm 1 as well as the discounted mean reward from the other SBSs at the $(t-1)$-th time slot, i.e., $\{\bar{\boldsymbol{\mu}}_{b'b}^{t-1}\}$.

C. Signalling Overhead and Computational Complexity Analysis

It is worth noticing that from the perspective of SBS b, the only information that need to be shared with SBS $b', \forall b' \in \mathcal{L}_b \backslash b$ at the end of time slot t for the proposed algorithm is its actual action decision, i.e., a scalar value a_b^t, and its discounted mean reward information, i.e., a vector $\bar{\boldsymbol{\mu}}_{bb'}^t$ with A entries. The resulting overall signalling overhead for the entire small cell network at each time slot is $O\left(N_b(N_b-1)(A+1)\right)$. On the contrary, the conventional distributed power allocation designs that are based on iterative inter-SBS fronthaul information exchange [15]–[20] require each SBS to exchange N_b positive integer scalars of key ICI coupling parameters with others in each iteration, resulting in an overall per-time slot inter-SBS communication overhead of $O\left(\xi N_b^2(N_b-1)\right)$, where ξ is the total number of iterations within a time slot and a typical ξ for convergence in a 3-SBS network is $\xi = 10$ [19]. Hence, the ratio of signalling overhead required by the proposed design to that of the conventional iterative distributed designs for the considered scenario is given by $\omega = \frac{A+1}{\xi N_b}$. It can be further reduced to $\omega = \frac{1}{\xi N_b}$ at some certain time slots by exchanging the reward information, i.e., $\bar{\boldsymbol{\mu}}_{bb'}^t$ of A entries, with SBS $b', b' \in \mathcal{L}_b \backslash b$ at a much lower frequency with a reasonable accuracy. As will be shown in Section IV, the reward information can be exchanged at a lower frequency without affecting much the accuracy. Thus, the proposed algorithm consumes lighter signalling overhead as compared to that of the conventional iterative distributed power allocation designs, especially for future dense networks, which can also be interpreted as the improvement of network scalability and latency suffering from the capacity-constrained fronthaul links.

The computational complexity of the proposed distributed design mainly depends on the calibrated forecasting part at the individual SBSs, i.e., Algorithm 1, whose complexity is of the order of $(N_b-1)N_\epsilon^{(A-1)}$, which implies a linearly dependence in N_b, a polynomial dependence in N_ϵ and an exponential dependence in A. Furthermore, the major complexity of Algorithm 1 comes from the calculation of $\{\boldsymbol{\psi}_{bb'}^{t*}\}$ over $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ for other SBSs $b', \forall b' \in \mathcal{L}_b \backslash b$ as per

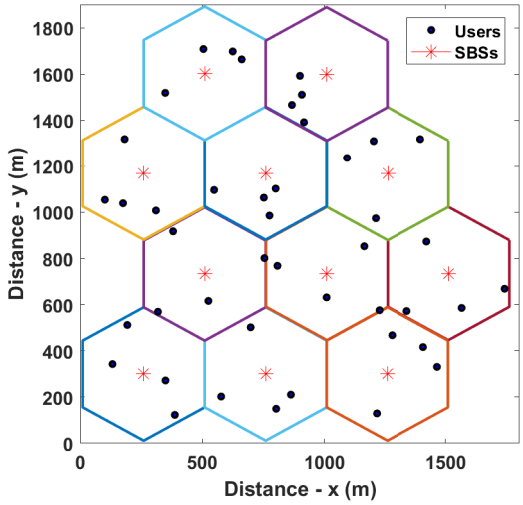


Fig. 2. An illustration of the simulation topology.

(12). Instead of solving exactly the min-max problem in (12) via linear programming, we can use the multiplicative weights algorithm [31] that allows a small violation $\delta > 0$ in each of A constraints given by (12), to approximate $\{\boldsymbol{\psi}_{bb'}^{t*}\}$, thus reduce significantly the complexity to $O(\frac{AN_\epsilon(N_b-1)}{\delta^2}\ln(N_\epsilon))$ [29]. The computational complexity can be further reduced by adopting a smaller value of N_ϵ with a reasonable accuracy.

IV. SIMULATION RESULTS

As shown in Fig. 2, we consider a distributed small cell network consisting of 11 adjacent SBSs, where we only focus on the top $N_b = 5$ neighboring interferers among all the SBSs. $N_u = 4$ users are randomly dropped in the vicinity of the boundary of each small cell to account for the worst-case ICI effect. The channel gain $\Psi_{bb}^t,\ b \in \mathcal{L}_b$, varies across individual time slots and the large-scale fading component is modeled by $\alpha_{bb}^t = G_b L_{bb}^t e^{-0.5\frac{(\sigma_s \ln 10)^2}{100}}$, where $G_b = 15$ dBi denotes the antenna gain, the path loss over a distance of l km is modeled as $L_{bb}^t = 128.1 + 37.6\log_{10}(l)$ [33], and $\sigma_s = 8$ dB represents the log-normal shadowing standard deviation. Other simulation parameters, unless otherwise stated, are described as follows, the AWGN variance

$\sigma_b^2 = -96$ dBm and the maximum Doppler frequency $f_d = 10$ Hz are identical to all users, the maximum transmit power is $P_b^{\max} = E_A = 3$ W [33] and the slot duration is $\tau = 20$ ms [27]. It has been proved in [26] and [34] that the binary power allocation achieves considerably good performance for sum-rate maximization problem. For the purpose of illustration, let us consider the scenario of $N_b = 5$ agents, i.e., 5 top interfering SBSs, with $A = 2$ discrete transmit power levels. Unless otherwise stated, transmit power levels 1 and 2 of the individual SBSs are set to be 0 W and 3 W, respectively, i.e., $\mathcal{E} = \{0, 3\}$ W, where $N_\epsilon = 6$ candidate probability values over $\mathcal{A}$ are set to be $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\} = \{(0,1), (0.2,0.8), (0.4,0.6), (0.6,0.4), (0.8,0.2), (1,0)\}$.

In order to demonstrate the advantages of our proposed distributed power allocation design, a non-cooperative power allocation design where each SBS autonomously and greedily chooses the transmit power from $\mathcal{E}$ to maximize its own performance without concerning its impact on others, has been set as the benchmark scheme. A random power allocation design that randomly selects the transmit power from $\mathcal{E}$ at the individual time slots has been employed to indicate the performance lower bound. Furthermore, an optimal centralized power allocation design that has full access to CSI and exhaustively searches over all combinations of power levels of the SBSs for the best transmit power decisions at each time slot at the cost of centralized coordination, is chosen as the performance upper bound. Note that a performance gap between the centralized and the distributed algorithms is expected, and our main focus is how much and how fast this gap can be closed using a distributed algorithm. For fair comparison, identical constraints have been applied to all power allocation designs and the channel condition varies across each individual time slot. Unless otherwise stated, each average simulation result for the individual power allocation designs is obtained by averaging all of the actual experimental points at the individual time slots over the previous 150 time slots.

A. Maximization of sum data rate

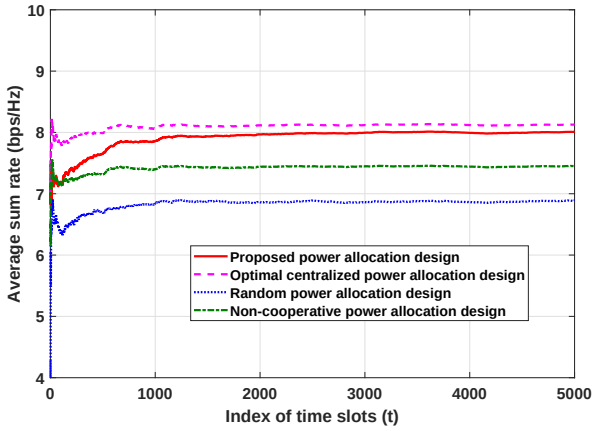
In this case study, we aim at maximizing the long-term overall sum-rate among all SBSs. Due to the fact that the users within the same small cell are served sequentially in their respective time slots in the considered scenario, the objective function in (5) can be expressed as

$$\max_{\{P_b^t\}} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{b=1}^{N_b} \left(R_b^t\right) \right\}. \tag{13}$$

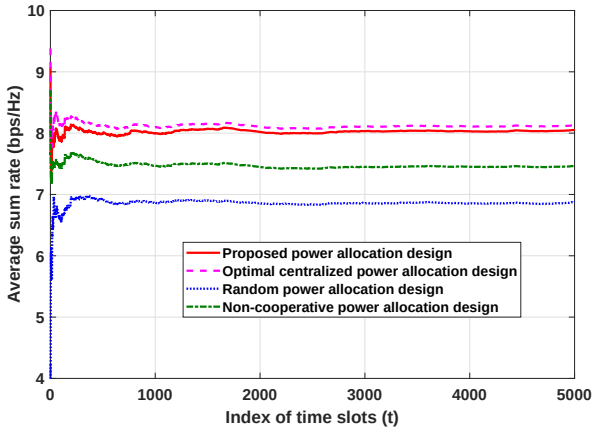
Note that we have further adapted the step 8 of Algorithm 2 such that the selected transmit power level is associated with the highest overall reward corresponding to the maximum estimated sum-rate, i.e., $a_b^t = \underset{a\in\mathcal{A}}{\operatorname{argmax}}(\hat{\boldsymbol{\mu}}_b^t + \sum_{b'=1,b'\neq b}^{N_b} \bar{\boldsymbol{\mu}}_{b'b}^{t-1})$.

Fig. 3 depicts the average sum-rate of various power allocation designs for both the highly uncertain environment, i.e., $f_d = 15$ Hz in Fig. 3(a) and $f_d = 10$ Hz in Fig. 3(b), and the near-static environment, i.e., $f_d \approx 0$ Hz in Fig. 3(c). For the purpose of better illustration, the performance, i.e., the average sum-rate, of all power allocation designs at individual time slots is obtained by averaging over all past time slots. As can be seen from the figure, the performance gap between the proposed distributed design and the optimal centralized design is narrowed with increasing time for both cases of $f_d = 10$ Hz and $f_d = 15$ Hz, whilst the performance of the proposed distributed design in Fig. 3(a) slightly degrades as compared to that of Fig. 3(b). This is due to the fact that instead of tracking the instantaneous channel conditions at the individual time slots, the proposed design adaptively tracks the recent trend of the wireless channel variations via gradually improving the predictions of the individual SBSs on others, and asymptotically optimizes the transmit power decisions at the SBSs.

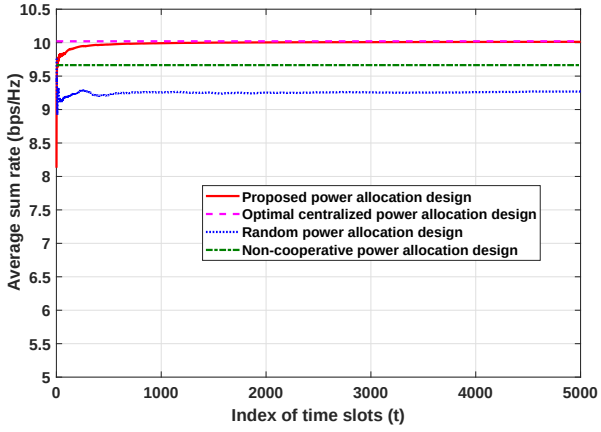
In order to better illustrate the performance of our proposed design, we select SBS 2 and SBS 4 as two representatives to show the evolution of their power decisions and the resulting average data rates. Fig. 4 illustrates the average data rate of SBS 2 and SBS 4 for various power allocation designs, respectively, for the near-static environment. As can be concluded from the figure, the proposed distributed design rapidly approaches towards the optimal centralized design



(a)

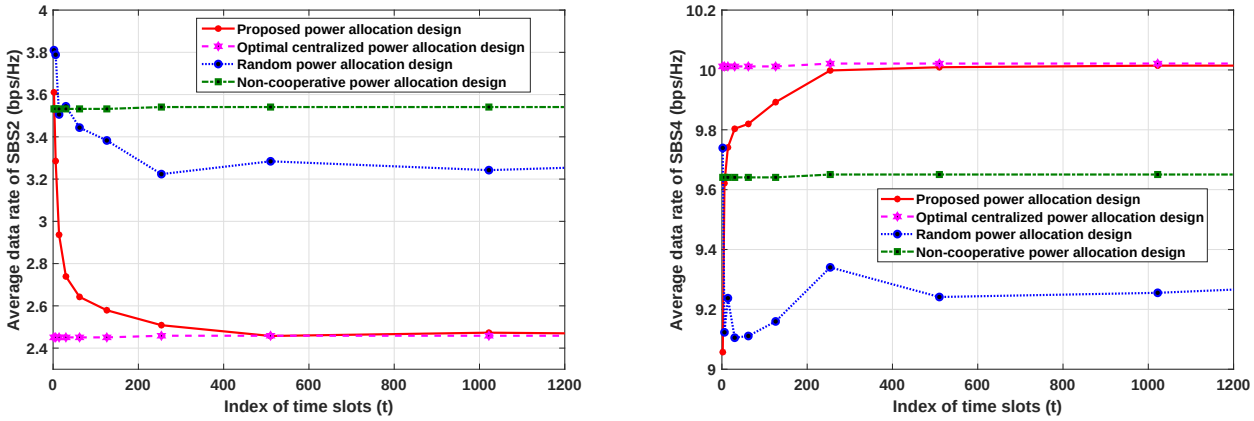


(b)



(c)

Fig. 3. Sum-rate of various power allocation designs for (a) $f_d = 15$ Hz, (b) $f_d = 10$ Hz, (c) $f_d \approx 0$ Hz.



(a) (b)

Fig. 4. Average data rate of (a) SBS2, (b) SBS4 for various power allocation designs.

within the first 200 time slots and follows it closely from the 500^{th} time slot onwards, for both SBS 2 and SBS 4. In comparison with Fig. 3(c), one may conclude that although choosing the maximum transmit power yields higher average data rate for SBS 2, the proposed design encourages the SBSs to select proper actions such that the system-level long-term sum data rate among all SBSs is maximized. Furthermore, Fig. 4 indicates that the non-cooperative power allocation design is not suitable for the optimization of the system-level performance as the individual SBSs aggressively choose the maximum transmit power regardless of others.

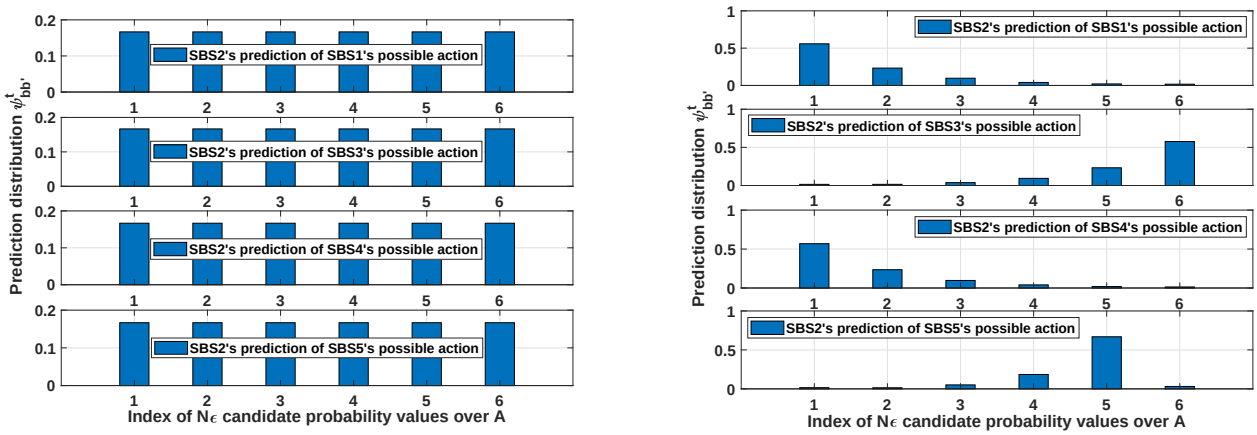
Table II depicts the possible sum data rate that could be achieved for various combinations of transmit power levels for the SBSs at the final time slot. One may observe from the table that the maximization of sum data rate of 10.0206 bps/Hz is achieved when the SBSs transmit, respectively, at power levels 2, 1, 1, 2 and 1.

Fig. 5 describes the prediction distribution of SBS 2, i.e., $\{\boldsymbol{\psi}_{2b'}^t\}$, over candidate probability values $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\} = \{(0,1), \ldots, (1,0)\}$ against SBS 1, SBS 3, SBS 4 and SBS 5 at the initial and the final time slots, respectively. Fig. 5(a) illustrates the uniformly distributed $\{\boldsymbol{\psi}_{2b'}^t\}$ at the initial time slot, where $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\}$ have equal likelihood to occur. Fig. 5(b) shows the

TABLE II

REWARD TABLE FOR SBSS AT FINAL TIME SLOT

Action combinations	(1,1,1,1,1)	(1,1,1,1,2)	(1,1,1,2,1)	(1,1,1,2,2)	(1,1,2,1,1)	(1,1,2,1,2)	(1,1,2,2,1)	(1,1,2,2,2)
Sum-rate (bps/Hz)	8.4554	8.4579	10.0197	10.0194	8.3360	8.3390	9.9015	9.9013
Action combinations	(1,2,1,1,1)	(1,2,1,1,2)	(1,2,1,2,1)	(1,2,1,2,2)	(1,2,2,1,1)	(1,2,2,1,2)	(1,2,2,2,1)	(1,2,2,2,2)
Sum-rate (bps/Hz)	8.2374	8.2387	9.7643	9.7636	8.1297	8.1320	9.6634	9.6631
Action combinations	(2,1,1,1,1)	(2,1,1,1,2)	(2,1,1,2,1)	(2,1,1,2,2)	(2,1,2,1,1)	(2,1,2,1,2)	(2,1,2,2,1)	(2,1,2,2,2)
Sum-rate (bps/Hz)	8.4588	8.4613	10.0206	10.0201	8.3399	8.3428	9.9025	9.9023
Action combinations	(2,2,1,1,1)	(2,2,1,1,2)	(2,2,1,2,1)	(2,2,1,2,2)	(2,2,2,1,1)	(2,2,2,1,2)	(2,2,2,2,1)	(2,2,2,2,2)
Sum-rate (bps/Hz)	8.2409	8.2422	9.7652	9.7646	8.1338	8.1360	9.6646	9.6642



(a) (b)

Fig. 5. Prediction distribution of SBS 2 of $N_\epsilon = 6$ candidate probability values over $\mathcal{A}$ at (a) the initial time slot, and (b) the final time slot.

optimal prediction distribution $\{\boldsymbol{\psi}^{t*}_{2b'}\}$ after the convergence. To be specific, SBS 2's prediction distribution over $\{\mathbf{p}_1, \ldots, \mathbf{p}_{N_\epsilon}\} = \{(0,1),(0.2,0.8),(0.4,0.6),(0.6,0.4),(0.8,0.2),(1,0)\}$ for SBS 5 is $\boldsymbol{\psi}^{t*}_{25} = [0.007, 0.019, 0.042, 0.107, 0.770, 0.055]$. In the light of Algorithm 1, SBS 2 will output its forecast result $\mathbf{p}^t_{25}$ of SBS 5 according to $\boldsymbol{\psi}^{t*}_{25}$. For instance, with the probability

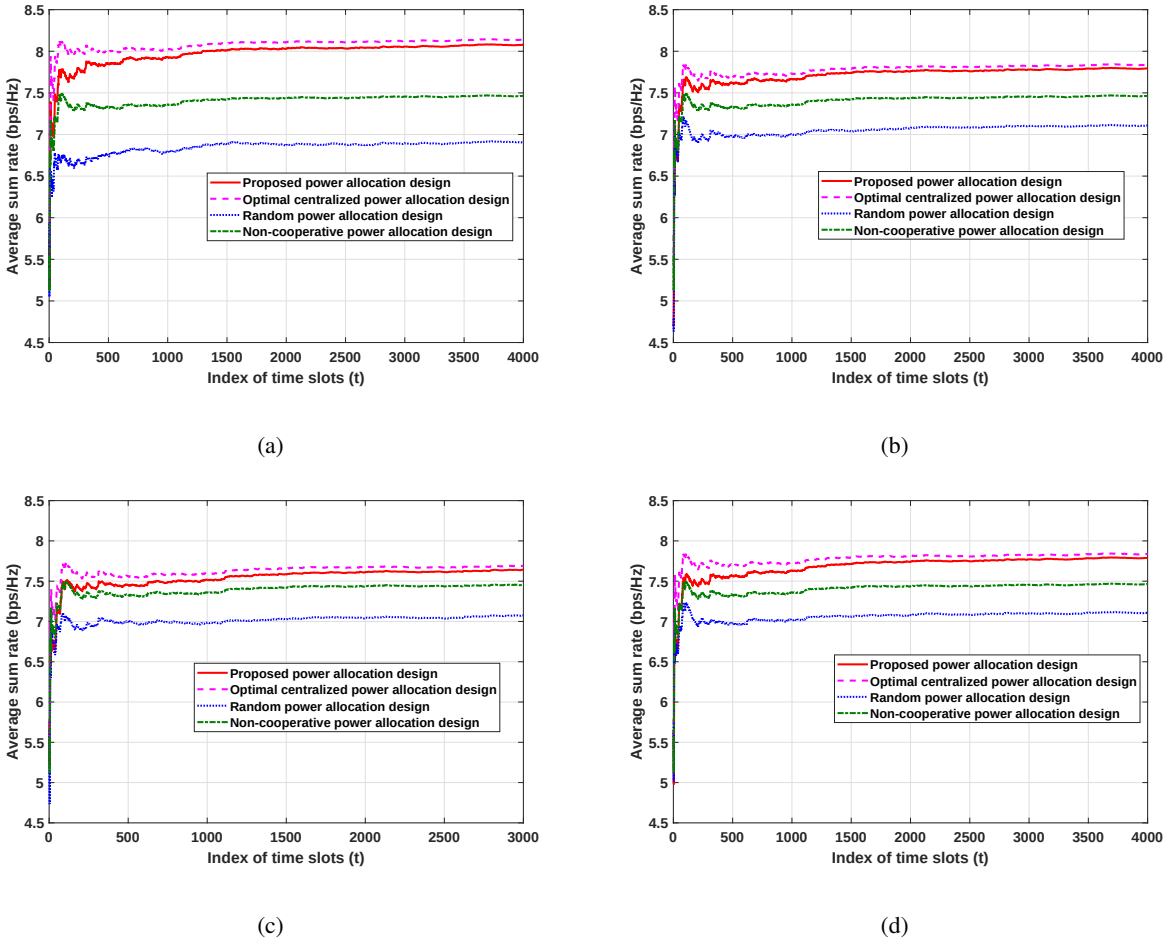


Fig. 6. Sum-rate of various power allocation designs for (a) 5 SBSs with $\mathcal{E} = \{0, 3\}$ W, (b) 5 SBSs with $\mathcal{E} = \{1, 3\}$ W, (c) 4 SBSs with $\mathcal{E} = \{1, 3\}$ W, (d) 5 SBSs with $\mathcal{E} = \{1, 3\}$ W and less frequent reward information exchange at every 5 time slots.

of 0.770, SBS 2 will forecast $\mathbf{p}_{25}^{t} = [0.8, 0.2]$ over SBS 5's possible transmit power levels 1 and 2, respectively, indicating that SBS 2's prediction of SBS 5 is highly likely to be power level 1. Then it is obvious from the figure that the SBSs 1, 3, 4 and 5 are highly likely to be expected by SBS 2 to choose actions 2, 1, 2 and 1, respectively. This conclusion is in agreement with Table II, where the individual SBSs selecting the combinations of transmit power levels of $(2, 1, 1, 2, 1)$ lead to the maximization of sum rate.

Fig. 6 shows the comparison results of our proposed design against various power allocation

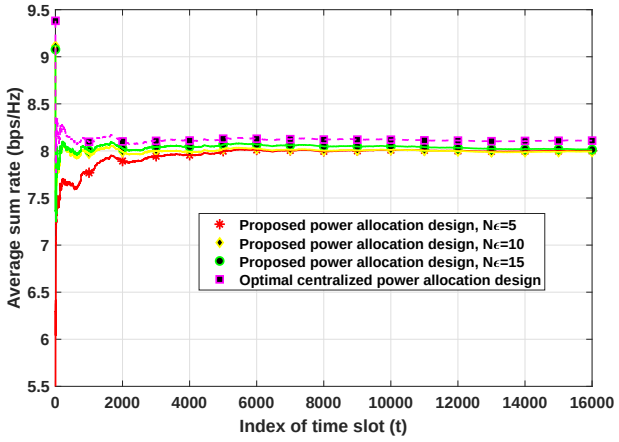


Fig. 7. Average sum rate for various number of N_ϵ candidate probability values for the proposed design for $f_d = 15$ Hz.

designs for different system settings. For the purpose of better illustration, the average sum-rate is obtained by averaging over all past time slots. Fig. 6(a) and Fig. 6(b) indicate the sum-rate performance for different transmit power levels, whilst Fig. 6(b) and Fig. 6(c) describe the sum-rate performance when different numbers of interfering SBSs are considered. One may conclude from the figures that binary power control, i.e., $\mathcal{E} = \{0, 3\}$ W, achieves better average sum-rate performance for both optimal centralized design and our proposed distributed design. Furthermore, the sum-rate performance degrades when smaller number of interfering SBSs is considered. This is due to the fact that though SBS 5 is not participating in the power control, it still poses interference on users associated to the other SBSs. As can be observed from Fig. 6(b) and Fig. 6(d), though exchanging reward information at a lower frequency may decrease the convergence speed, the accuracy remains similar to the case where the SBSs exchange reward information at the end of every time slot.

Fig. 7 shows the average sum rate for various number of N_ϵ, i.e., number of candidate probability values over $\mathcal{A}$, for the proposed design for $f_d = 15$ Hz. For the purpose of better illustration, the average sum-rate is obtained by averaging over all past time slots. As can be

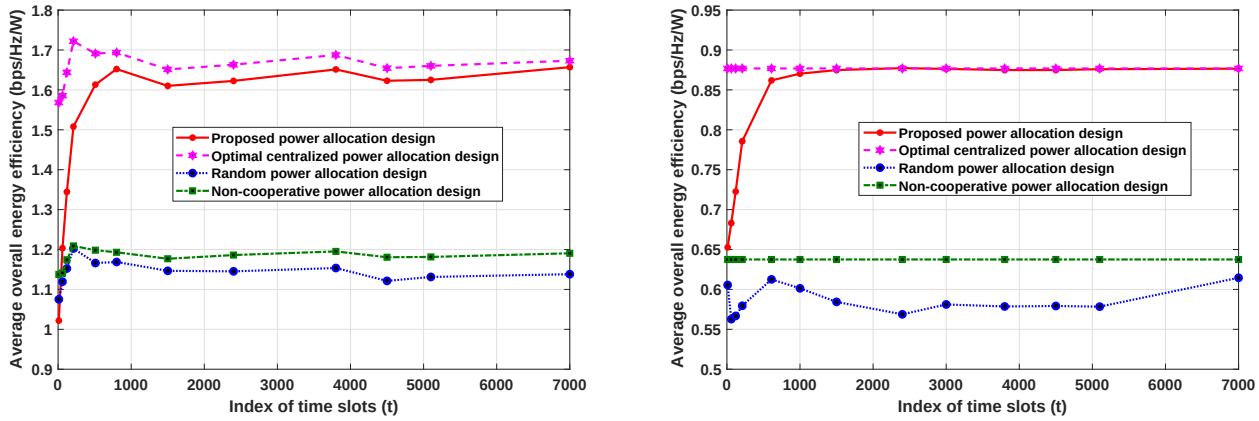


Fig. 8. Overall energy efficiency of various power allocation designs for (a) $f_d = 10$ Hz, (b) $f_d \approx 0$ Hz.

observed from the figure, the performance of the proposed design for various number of N_ϵ approaches towards roughly the same point. However, the convergence speed is significantly improved when the number of N_ϵ increases from 5 to 10, whilst minor differences can be observed for the performance of $N_\epsilon = 10$ and $N_\epsilon = 15$. Therefore, the computational complexity the our proposed design can be further reduced by selecting a smaller value of N_ϵ with a reasonable convergence speed and accuracy.

B. Maximization of overall energy efficiency

Next, we consider the problem of maximizing the long-term overall energy efficiency among all SBSs, and convert the objective function in (5) as

$$\max_{\{P_b^t\}} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{b=1}^{N_b} R_b^t}{\sum_{b=1}^{N_b} P_b^{\mathrm{t}} + P_c} \right\}, \tag{14}$$

where $P_c = 2$ W is the total hardware circuit power consumption of the network. Furthermore, the step 8 of Algorithm 2 is adapted as $a_b^t = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\left(\frac{\hat{\boldsymbol{\mu}}_b^t + \sum_{b'=1,b'\neq b}^{N_b} \bar{\boldsymbol{\mu}}_{b'b}^{t-1}}{P_b^t + \sum_{b'=1,b'\neq b}^{N_b} \sum_{a=1}^{A} p_a^t E_a + P_c}\right)$, to select the transmit power level associated with the highest estimated overall energy efficiency.

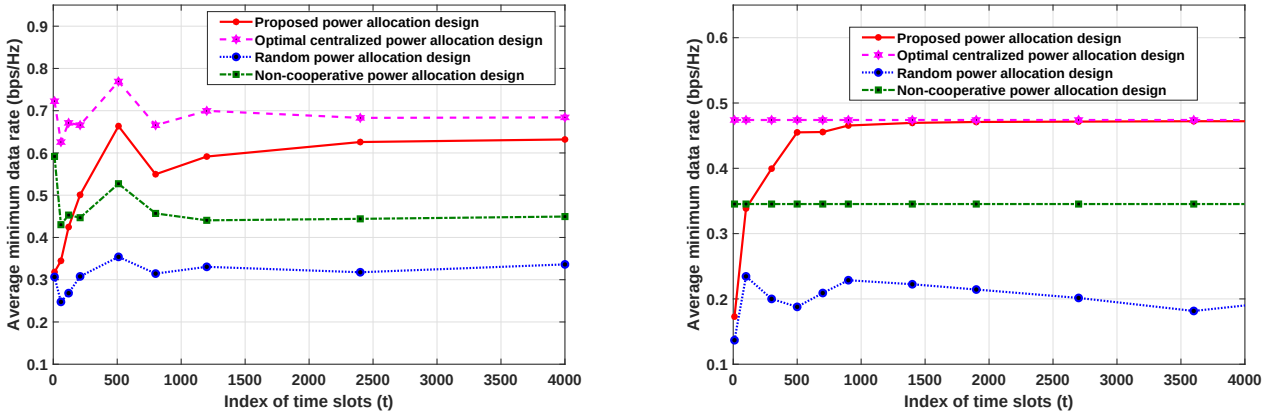
Fig. 8 presents the average overall energy efficiency of various power allocation designs for the time-varying wireless channel conditions in Fig. 8(a), and the near-static environment in Fig. 8(b), respectively. One may observe from Fig. 8 that the proposed distributed design closely follows the optimal centralized design from approximately the 600^{th} time slot onwards for the case of $f_d \approx 0$ Hz, whilst for the case of uncertain channel variations, the performance gap between the proposed design and the centralized design decreases with increasing time, which demonstrates the improvement of forecast accuracy at the individual SBSs.

C. Maximization of minimum achievable data rate

Finally, we consider the problem of maximizing the minimum achievable data rate among all SBSs. For the purpose of better illustration, let us set $\mathcal{E} = \{1, 3\}$ W and recast the objective function in (5) as

$$\max_{\{P_b^t\}} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \min_{b} \left(R_b^t\right) \right\}. \tag{15}$$

We further adapt the step 8 of Algorithm 2 as $a_b^t = \underset{a\in\mathcal{A}}{\operatorname{argmax}} \min([\hat{\boldsymbol{\mu}}_b^t | \{\bar{\boldsymbol{\mu}}_{b'b}^{t-1}\}])$, to select the transmit power level associated with the highest estimated overall max-min reward.



(a) (b)

Fig. 9. Minimum data rate of various power allocation designs for (a) $f_d = 10$ Hz, (b) $f_d \approx 0$ Hz.

Fig. 9 presents the average minimum data rate for various power allocation designs for the time-varying wireless channels in Fig. 9(a), and the near-static channel conditions in Fig. 9(b), respectively. As can be seen from Fig. 9, the proposed design sharply approaches towards the optimal centralized design for the case of near-static environment within approximately 500 time slots and has a slightly degraded performance for the case of highly uncertain environment, whilst the performance gap between the proposed design and the centralized design is narrowed with increasing number of time slots. In addition, the proposed design outperforms the benchmark scheme in both cases. This is due to the fact that in the benchmark scheme, each individual SBS autonomously selects its transmit power level without taking into account either the impact it may have on its counterparts or the variations in wireless channel conditions.

V. CONCLUSION

This paper proposes a combined calibrated learning and bandit approach to online transmit power control in distributed small cell networks operated under a single frequency band, which asymptotically maximizes the average reward in the long run. The proposed design allows the individual SBSs to gradually improve their predictions on the behaviour of the other SBSs, and react with the best response based on the forecasted results as well as the past observations at the individual time slots. Numerical simulation results validate that the proposed distributed design outperforms the benchmark scheme and closely follows the optimal centralized design with limited amount of information exchange for all case studies.

REFERENCES

[1] Cisco, “Cisco Visual Networking Index: Forecast and Methodology, 2016-2021”, White Paper, Sep. 2017.

[2] F. Al-Turjman, E. Ever and H. Zahmatkesh, “Small Cells in the Forthcoming 5G/IoT: Traffic Modelling and Deployment Overview,” in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 28-65, Firstquarter 2019.

[3] J. Wang, W. Guan, Y. Huang, R. Schober and X. You, “Distributed Optimization of Hierarchical Small Cell Networks: A GNEP Framework,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 2, pp. 249-264, Feb. 2017.

[4] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella and D. Wbben, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68-76, May 2014.

[5] K. Mun, "CBRS: New Shared Spectrum Enables Flexible Indoor and Outdoor Mobile Solutions and New Business Models", White paper, Mar. 2017.

[6] MulteFire Alliance, , "MulteFire release 1.0 technical paper", White paper, Jan. 2017.

[7] FCC, "FCC Rule Making on 3.5 GHz Band/Citizens broadband radio service", Apr. 2015.

[8] J. Xu, J. Wang, Y. Zhu, Y. Yang, X. Zheng, S. Wang, L. Liu, K. Horneman and Y. Teng, "Cooperative Distributed Optimization for the Hyper-Dense Small Cell Deployment," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 61-67, May 2014.

[9] A. Fehske, I. Viering, J. Voigt, E. Sartori, S. Redana and G. P. Fettweis, "Small-cell self-organizing wireless networks ", *Proceedings of the IEEE*, vol. 102, no. 3, pp. 334350, Mar. 2014.

[10] X. Zhang, M. R. Nakhai, G. Zheng, S. Lambotharan and B. Ottersten, "A Calibrated Learning Approach to Distributed Power Allocation in Small Cell Networks", accepted by ICASSP 2019.

[11] T. A. Le and M. R. Nakhai, "Downlink Optimization with Interference Pricing and Statistical CSI," in *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2339-2349, Jun. 2013.

[12] Z. Han and K. J. R. Liu, "Noncooperative Power-Control Game and Throughput Game over Wireless Networks," in *IEEE Transactions on Communications*, vol. 53, no.10 pp. 1625-1629, Oct. 2005.

[13] G. Bacci, E. V. Belmega, P. Mertikopoulos, L. Sanguinetti, " Energy-Aware Competitive Power Allocation for Heterogeneous Networks Under QoS Constraints," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4728-4742, Sept. 2015.

[14] Z. Wang, B. Hu, X. Wang and S. Chen, "Interference Pricing in 5G Ultra-Dense Small Cell Networks: A Stackelberg Game Approach," *IET Communications*, vol. 10, no. 15, pp. 1865-1872, Oct. 2016.

[15] Z. Xiang, M. Tao and X. Wang, "Coordinated Beamforming Design in Multicell Multicast Networks," in *IEEE Transactions on Wireless Communications*, vol. 12, pp. 12-21, Jan. 2013.

[16] X. Zhang and M. R. Nakhai, "Robust Chance-Constrained Distributed Beamforming for Multicell Interference Networks," in *IEEE International Conference on Communications (ICC)*, May 2016.

[17] A. Shaverdian and M. R. Nakhai, "Robust Distributed Beamforming with Interference Coordination in Downlink Cellular Networks," in *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2411-2421, Jul. 2014.

[18] H. Pennanen, A. Tolli, and M. Latva-aho, "Decentralized Robust Beamforming for Coordinated Multi-Cell MISO Networks," in *IEEE Signal Processing Letters*, vol. 21, no. 3, pp. 334-338, Mar. 2014.

[19] O. Tervo, H. Pennanen, D. Christopoulos, S. Chatzinotas and B. Ottersten, "Distributed Optimization for Coordinated Beamforming in Multicell Multigroup Multicast Systems Power Minimization and SINR Balancing," in *IEEE Transactions on Signal Processing*, vol. 66, no. 1, pp. 171-185, Jan. 2018.

[20] C. Shen, T. H. Chang, K. Y. Wang, Z. Qiu, and C. Y. Chi, "Distributed Robust Multicell Coordinated Beamforming With Imperfect CSI: An ADMM Approach," in *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2988-3003, Jun. 2012.

[21] S. Maghsudi and S. Stanczak, "Channel Selection for Network-Assisted D2D Communication via No-Regret Bandit Learning With Calibrated Forecasting," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309-1322, Mar. 2015.

[22] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan and D. Matolak, "A Machine Learning Approach for Power Allocation in HetNets Considering QoS," *IEEE ICC*, pp. 1-7, May 2018.

[23] Y. S. Nasir and D. Guo, "Deep Reinforcement Learning for Distributed Dynamic Power Allocation in Wireless Networks," in *arXiv:1808.00490*.

[24] F. Meng, P. Chen and L. Wu, "Power Allocation in Multi-User Cellular Networks With Deep Q Learning Approach," in *arXiv:1812.02979*.

[25] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu and N. D. Sidiropoulos, "Learning to Optimize: Training Deep Neural Networks for Wireless Resource Management," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5438-5453, Oct. 2018.

[26] Y. Shen, Y. Shi, J. Zhang and K. B. Letaief, "LORA: Learning to Optimize for Resource Allocation in Wireless Networks with Few Training Samples," in *arXiv:1812.07998*.

[27] T. Kim, D. J. Love and B. Clerckx, "Does Frequent Low Resolution Feedback Outperform Infrequent High Resolution Feedback for Multiple Antenna Beamforming Systems?" *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1654-1669, Apr. 2011.

[28] N. Cesa-Bianchi and G. Lugosi, "Prediction, Learning, and Games," *Cambridge University Press*, Cambridge, UK, 2006.

[29] S. Mannor and G. Stoltz, "A Geometric Proof of Calibration," *Mathematics of Operations Research*, vol. 35, no. 4, pp. 721-727, Nov. 2010.

[30] D. Blackwell, "An Analog of the Minimax Theorem for Vector Payoffs," *Pacific Journal of Mathematics*, vol. 6, no. 1, pp. 1-8, 1956.

[31] Y. Freund and R. E. Schapire, "Adaptive Game Playing Using Multiplicative Weights," *Games and Economic Behavior*, vol. 29, pp. 79-103, 1999.

[32] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time Analysis of the Multi-armed Bandit Problem," *Machine learning*, vol. 47, no. 2-3, pp. 235-256, May 2002.

[33] 3GPP, "TR 36.814 V9.2.0: Further Advancements for E-UTRA Physical Layer Aspects (Release 9)," *Available online: http://www.3gpp.org*, Mar. 2017.

[34] J. Kim and D-H. Cho, "A Joint Power and Subchannel Allocation Scheme Maximizing System Capacity in Indoor Dense Mobile Communication Systems," in *IEEE Transactions on Vehicular Technology*, vol. 59, no. 9, pp. 4340-4353, Nov. 2010.